

S1 Brief Introduction of Cube Method

Balanced sampling aims to select a sample such that the Horvitz-Thompson estimator of auxiliary variables matches their known population totals exactly or approximately. This balance enhances the efficiency of estimators for variables of interest correlated with the auxiliary ones. The cube method [1] is a pioneering algorithm to achieve this goal. It consists of two sequential phases: the *flight phase* and the *landing phase*, which jointly ensure the sample's adherence to balance constraints while respecting predefined inclusion probabilities.

The Flight Phase. The flight phase iteratively transforms the initial inclusion probability vector into a sample by solving a system of balancing equations. At each step, it identifies a direction in the null space of the constraint matrix defined by auxiliary variable u , then updates the inclusion probabilities along this direction while maintaining their bounds between 0 and 1. This process continues until no further movement can be made without violating the constraints. The output is a random vector π with most components equal to 0 or 1, representing nearly finalized inclusions and exclusions, while a small subset of units may remain fractional.

The Landing Phase. In cases where undecided units persist after the flight phase, the landing phase deterministically selects a final sample to minimize the deviation from balance constraints. This is achieved by solving a discrete optimization problem over the remaining undecided units, choosing the combination that best satisfies the balancing conditions. If exact balance is infeasible, the algorithm prioritizes minimal imbalance through Euclidean distance minimization or other predefined metrics.

The cube method guarantees the selection of a well-balanced sample with design-unbiased estimators. Under regularity conditions, the resulting estimators exhibit asymptotic normality and enhanced statistical efficiency, as the enforced balance reduces variance compared to conventional sampling designs. This property holds even when the landing phase introduces negligible residual imbalance.

S2 Variance Estimation

The variance estimator for the simple random sampling labeling strategy using sample mean estimator ($\hat{\theta} = \frac{1}{n_b} \sum_{j=1}^{n_b} Y_j$, *classical*) holding budget $\pi_0 = \frac{n_b}{n}$ is

$$\text{Var}(\theta^*) = \sum_{i=1}^n \frac{(1 - \pi_0)\xi_i}{\pi_0^2} Y_i^2.$$

The variance estimator for prediction-powered inference with GD estimator [2, 3] using a uniform random labeling strategy ($\hat{\theta} = \frac{1}{n} \sum_{i=1}^n [\hat{f}(X_i) + (Y_i - \hat{f}(X_i)) \frac{\xi_i}{\pi_0}]$, *uniform*) holding budget $\pi_0 = \frac{n_b}{n}$ is

$$\text{Var}(\hat{\theta}) = \sum_{i=1}^n \frac{(1 - \pi_0)\xi_i}{\pi_0^2} [Y_i - f(X_i)]^2.$$

The variance estimator for traditional active inference based on independent sampling strategies designed using machine learning model predictions with GD estimator [2, 4] ($\hat{\theta} = \frac{1}{n} \sum_{i=1}^n [\hat{f}(X_i) + (Y_i - \hat{f}(X_i)) \frac{\xi_i}{\pi(X_i)}]$, *traditional-active*) is

$$\text{Var}(\hat{\theta}) = \sum_{i=1}^n \frac{(1 - \pi(X_i))\xi_i}{\pi^2(X_i)} [Y_i - f(X_i)]^2.$$

The variance estimator [5] for our method *cube-active* is

$$\text{Var}(\tilde{\theta}) = \sum_{i=1}^n \frac{(1 - \pi(X_i))\xi_i}{\pi^2(X_i)} [(Y_i - f(X_i)) - \beta \hat{u}(X_i)]^2,$$

where the β is a weighted least squares coefficient

$$\beta = \left(\sum_{i=1}^n c_i \frac{\hat{u}^2(X_i)}{\pi^2(X_i)} \right)^{-1} \sum_{i=1}^n c_i \frac{(Y_i - f(X_i)) \cdot \hat{u}(X_i)}{\pi^2(X_i)},$$

and c_i are the solutions of the nonlinear system

$$1 - \pi(X_i) = c_i - \frac{c_i \cdot \hat{u}(X_i)}{\pi(X_i)} \left(\sum_{\ell=1}^n c_\ell \frac{\hat{u}^2(X_\ell)}{\pi^2(X_\ell)} \right)^{-1} \frac{c_i \hat{u}(X_i)}{\pi(X_i)}.$$

S3 Additional Experimental Results

We present additional experimental results that could not be fully included in the main text due to space constraints. Specifically, we detail the outcomes from two more synthetic datasets, five more real-world regression datasets, and one binary classification dataset.

Linear (synthetic). The data generation process follows a linear regression scenario. The input features are independently and identically distributed as $X = (X^{(1)}, \dots, X^{(10)})^\top \in \mathbb{R}^{10}$, where each component $X^{(j)}$ is independently drawn from a standard normal distribution $\mathcal{N}(0, 1)$. The response variable Y is generated through the relationship:

$$Y = \sum_{j=1}^5 X^{(j)} + 0.3\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ represents independent Gaussian noise. The functional dependence involves linear combinations of the first five features, while the remaining five features $X^{(6)}, \dots, X^{(10)}$ are non-informative covariates that do not influence the response variable. All components of X and the noise variable ε are mutually independent. We generate 10,000 instances. Figure S1 demonstrates that our method achieves the lowest RMSE among all compared approaches, suggesting minimal estimation bias. The confidence intervals produced by our method are approximately 30% narrower than those from traditional active while maintaining coverage rates close to the nominal 90% level.

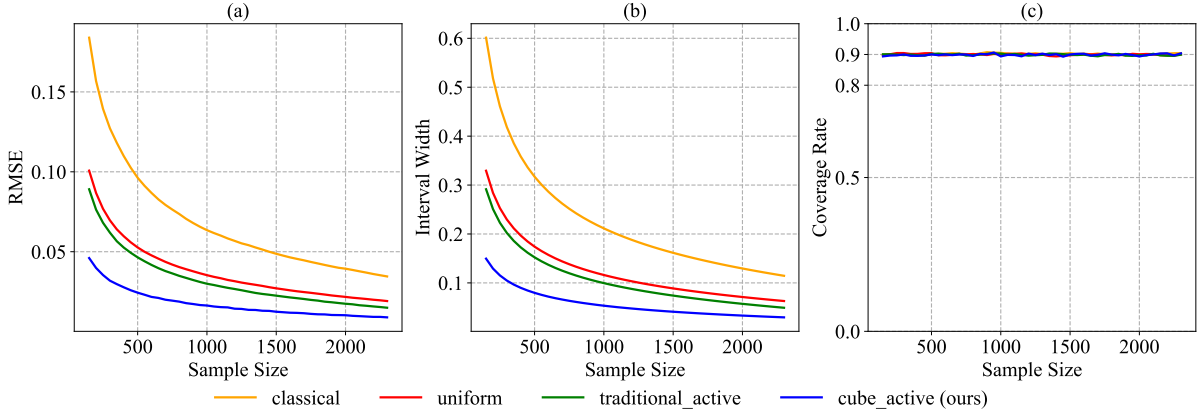


Figure S1: Performance comparison on the linear synthetic dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

Nonlinear (synthetic). The data generation process follows a nonlinear regression scenario. The input features are independently and identically distributed as $X = (X^{(1)}, \dots, X^{(15)})^\top \in \mathbb{R}^{15}$, where each component $X^{(j)}$ is independently drawn from a uniform distribution over $[-3, 3]$. The response variable Y is generated through the relationship:

$$Y = \sin(X^{(1)}) + 0.5(X^{(2)})^3 + 0.8 \exp(-|X^{(3)}|) + 0.6X^{(4)}X^{(5)} + 0.5\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ represents independent Gaussian noise. The functional dependence involves the first five features through distinct nonlinear transformations: a sinusoidal function of $X^{(1)}$, a cubic term of $X^{(2)}$, an exponentially decaying function of $|X^{(3)}|$, and an interaction term between $X^{(4)}$ and $X^{(5)}$. The remaining ten features $X^{(6)}, \dots, X^{(15)}$ are non-informative covariates. All components of X and the noise variable ε are mutually independent. We generate 10,000 instances. As shown in Figure S2, our method

consistently yields the lowest RMSE and narrowest confidence intervals across comparative methods, with coverage rates remaining stable around the nominal 90% threshold.

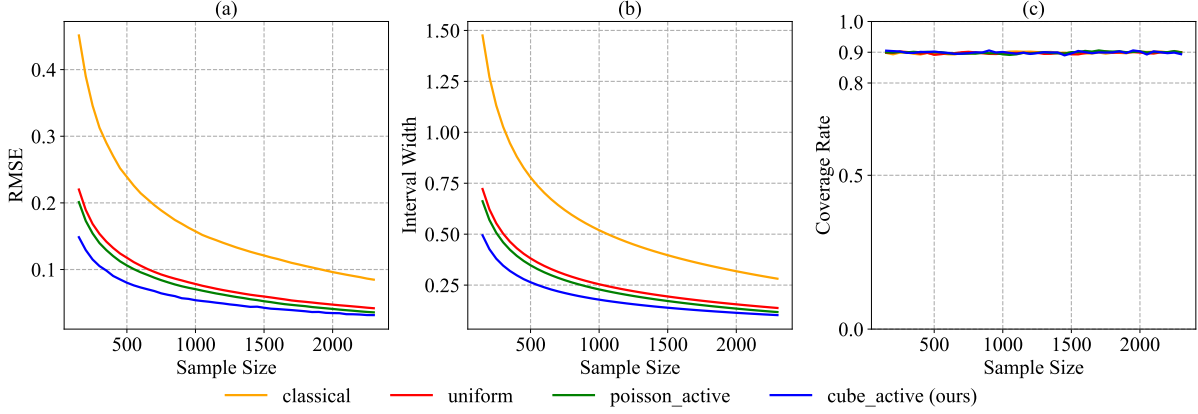


Figure S2: Performance comparison on the nonlinear synthetic dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

Communities and Crime. The Communities and Crime dataset from the UCI repository includes socio-economic and demographic information from various US communities, aiming to predict crime rates. The dataset comprises 1994 instances with 127 features, including attributes such as median income, population density, police presence, and unemployment rates [6, 7, 8]. The target variable is the total number of violent crimes per 100 000 population. Our method consistently provides superior performance by attaining the lowest RMSE and narrower confidence intervals compared to other methods, maintaining stable coverage around the nominal 90% confidence level.

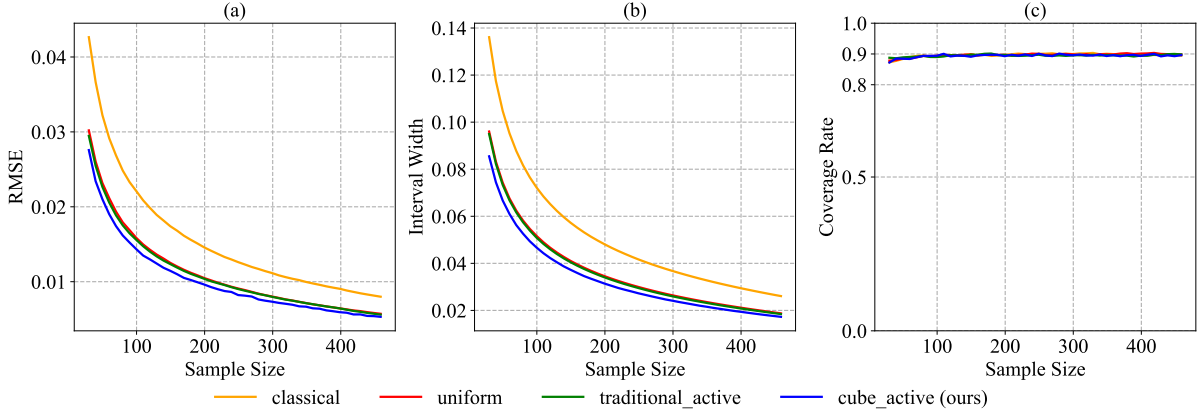


Figure S3: Performance comparison on the Communities and Crime dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

Concrete Compressive Strength. The Concrete Compressive Strength dataset from the UCI repository consists of 1030 instances and 8 quantitative input variables, including cement content, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age [9]. The target variable is concrete compressive strength measured in megapascals (MPa). Figure S4 shows our method consistently achieves the lowest RMSE and narrower confidence intervals compared to other methods, maintaining robust coverage around the nominal 90% confidence level.

Life Expectancy. The Life Expectancy dataset from Kaggle’s WHO database comprises 2938 instances with 20 socio-economic and health-related attributes such as GDP, adult mortality, alcohol consumption, vaccination coverage, and BMI [10]. The target variable is life expectancy at birth. Figure S5 illustrates

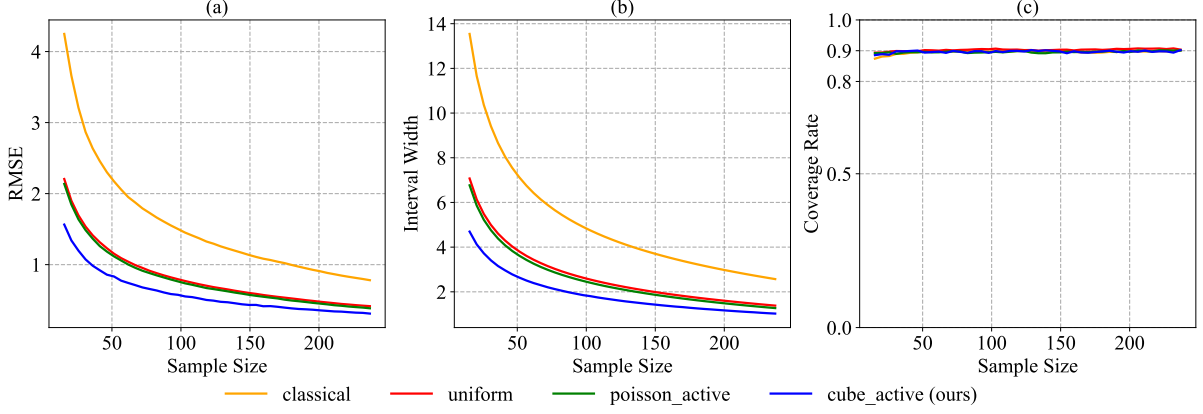


Figure S4: Performance comparison on the Concrete Compressive Strength dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

that our method consistently achieves the best predictive accuracy with the lowest RMSE, narrower confidence intervals, and robust empirical coverage around 90%.

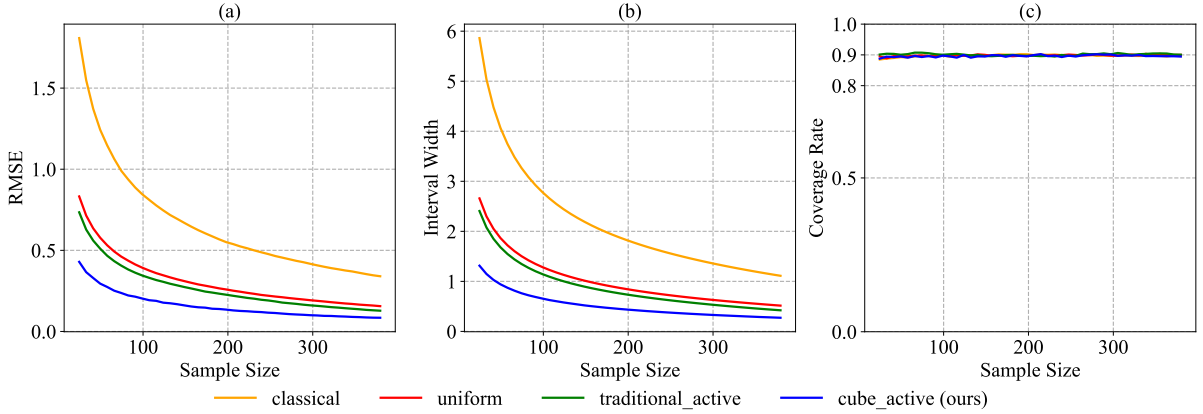


Figure S5: Performance comparison on the Life Expectancy dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

Superconductivity Data. The Superconductivity dataset from the UCI repository contains 21263 instances described by 81 physical and chemical attributes [11, 12, 8]. The dataset aims to predict the critical temperature at which materials exhibit superconductivity. Figure S6 indicates our method consistently provides superior results, achieving the lowest RMSE and narrower confidence intervals, while maintaining empirical coverage close to the nominal 90% level.

Post-election Survey Research. We further evaluate our method on a real-world survey dataset collected by the Pew Research Center after the 2020 U.S. presidential election [13]. The dataset contains 12,378 responses from a nationally representative sample, with binary labels indicating approval or disapproval of political messaging. Each response is associated with a range of demographic features, including age, gender, education level, and political affiliation. Figure S7 shows that while all three learning-based sampling strategies (uniform, traditional-active, and cube-active) perform comparably, cube-active still achieves a consistent 5% improvement in budget efficiency relative to traditional active inference across 10,000 repeated experiments, demonstrating the method’s robustness under practical constraints.

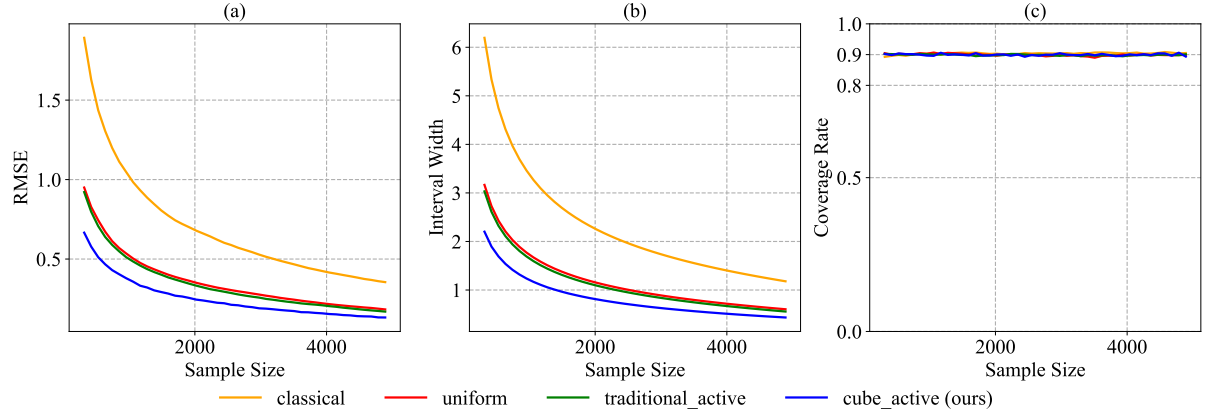


Figure S6: Performance comparison on the Superconductivity dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different methods.

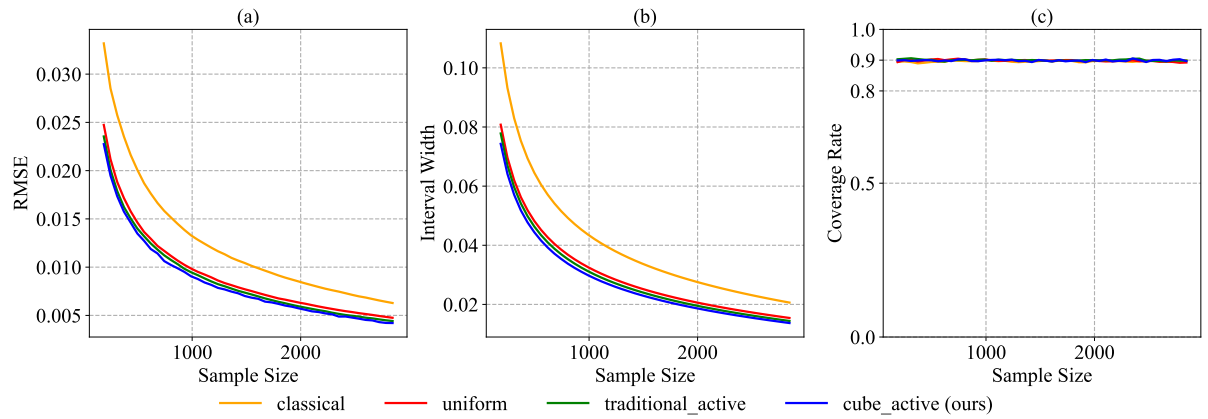


Figure S7: Performance comparison on the Post-election Survey dataset. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

S4 Experimental Details

This section provides supplementary details regarding our experimental setup.

Data Splits. All datasets underwent standardized preprocessing before being partitioned. Each dataset was equally divided into training and test sets using a 50%-50% split ratio. The training set served for model development, while the test set remained strictly isolated for final performance evaluation in our primary experiments.

Model Architectures and Hyperparameter Configurations. We implemented paired XGBoost regressors for both label prediction (\hat{f}) and uncertainty estimation (\hat{u}) across all experimental scenarios. Table S1 details the complete hyperparameter specifications per dataset. The detailed setting of parameters for each dataset is provided in Table S1, and the complete implementation leverages XGBoost’s optimized gradient boosting framework with early stopping criteria implicitly controlled through fixed estimator counts to prevent overfitting.

Table S1: Model Configurations and Hyperparameter Specifications

Dataset	Model	Estimators	Learning Rate	Max Depth	Objective Function
Credit Fraud	\hat{f}	2000	0.001	7	$-\sum [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$
Credit Fraud	\hat{u}	1000	0.001	7	$\sum (y_i - \hat{y}_i)^2$
Post-election	\hat{f}	1000	0.001	7	$-\sum [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$
Post-election	\hat{u}	1000	0.001	7	$\sum y_i - \hat{y}_i $
Other Datasets	\hat{f}, \hat{u}	1000	0.001	7	$\sum (y_i - \hat{y}_i)^2$

Evaluation Metrics. For each experimental configuration, we first partitioned the dataset into training and test subsets to train the estimator models $\hat{f}(\cdot)$ and $\hat{u}(\cdot)$. Subsequently, we conducted $T = 10,000$ independent Monte Carlo simulations for each method under comparison. Let θ denote the ground truth mean of the target variable in the test set. For each method and simulation $i \in \{1, \dots, T\}$, we computed the parameter estimate $\hat{\theta}_i$.

The root mean squared error (RMSE) for method j was then calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{T} \sum_{i=1}^T (\hat{\theta}_i - \theta)^2}, \quad (\text{S1})$$

providing a robust measure of estimation accuracy. Additionally, we computed the average confidence interval width at the 90% confidence level across all simulations and the average coverage of the real population mean θ .

S5 Sensitivity Analysis of τ

We present additional experimental results for the sensitivity analysis of τ on several datasets. The results are from Figure S8 to Figure S12 while keeping budget=0.1 fixed. All coverage rates in the above results closely approximate the target confidence level, confirming the validity of our findings.

Across all datasets, the RMSE and interval width generally maintained which in the main conclusion. At $\tau = 0$: The performance of *traditional_active* becomes nearly equivalent to *uniform*, as both rely solely on simple random sampling. As τ increases, both *uniform* and *traditional_active* exhibit precision improvement (decreasing RMSE and interval width), reaching a lower value around $\tau = 0.5$ for different datasets. This demonstrably quantifies the precision gains enabled by active sampling. When near $\tau = 1$, RMSE and interval width increase on some datasets. This occurs because the model for estimating prediction uncertainty (\hat{u}) is inherently imperfect. When the model erroneously assigns near-zero uncertainty ($\hat{u} \approx 0$) to specific data points, it can significantly inflate the estimator variance. These findings are well-aligned with conclusions established in those of prior work [4].

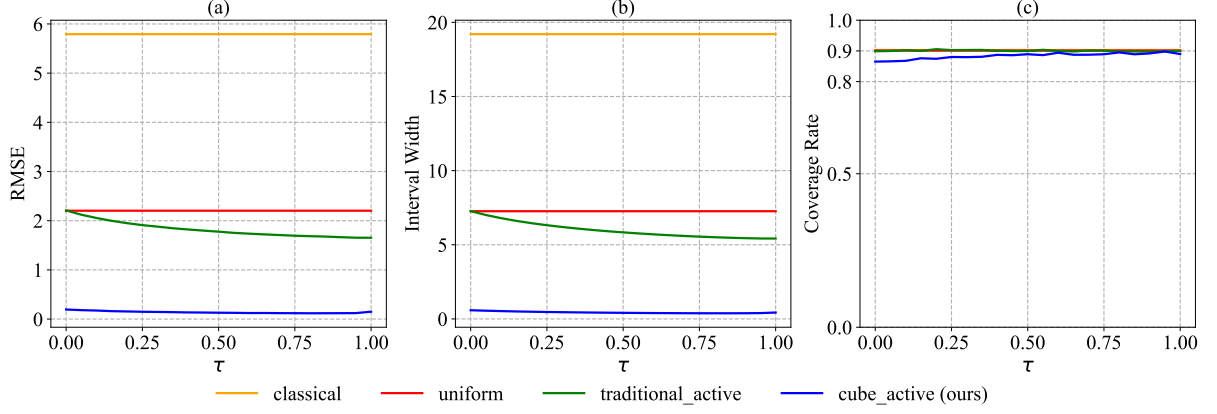


Figure S8: Performance comparison on the Bike Sharing dataset across different τ values while keeping budget=0.1 fixed. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

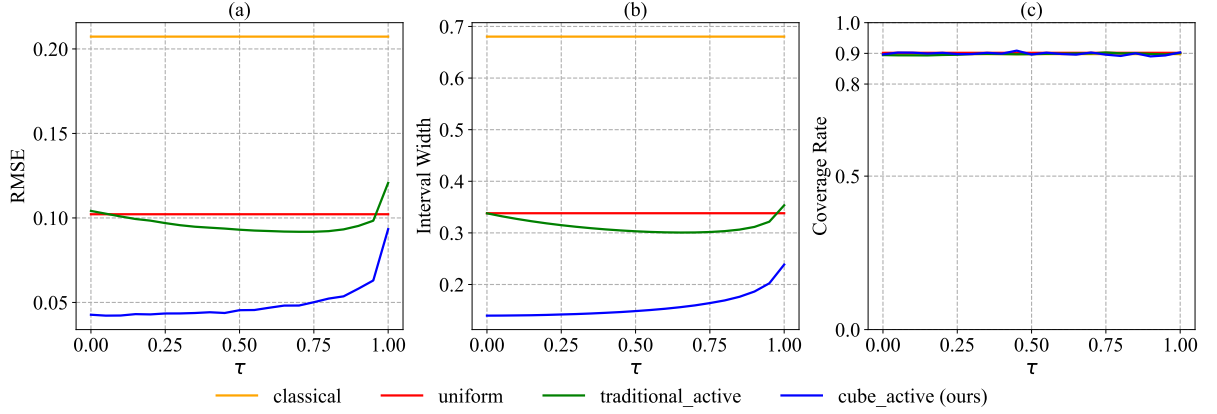


Figure S9: Performance comparison on the synthetic Friedman dataset across different τ values while keeping budget=0.1 fixed. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

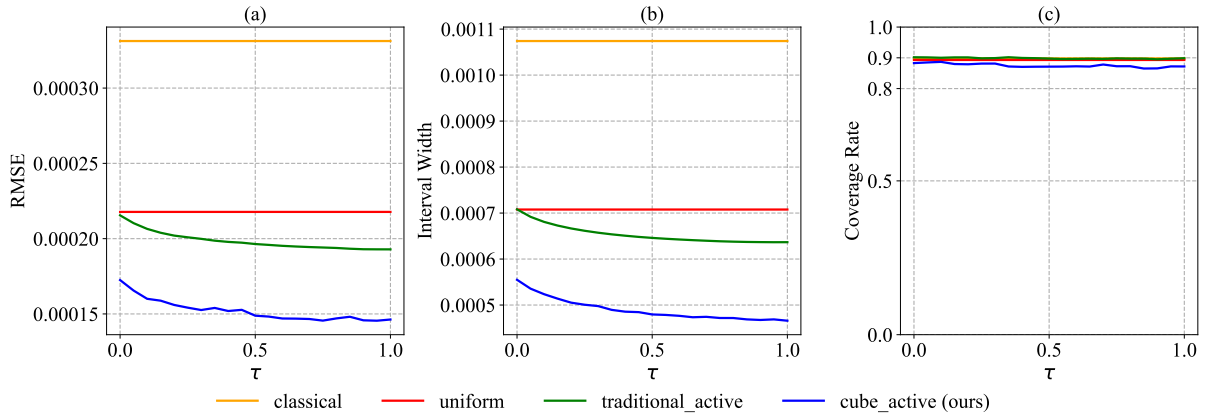


Figure S10: Performance comparison on the Credit Fraud Detection dataset across different τ values while keeping budget=0.1 fixed. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

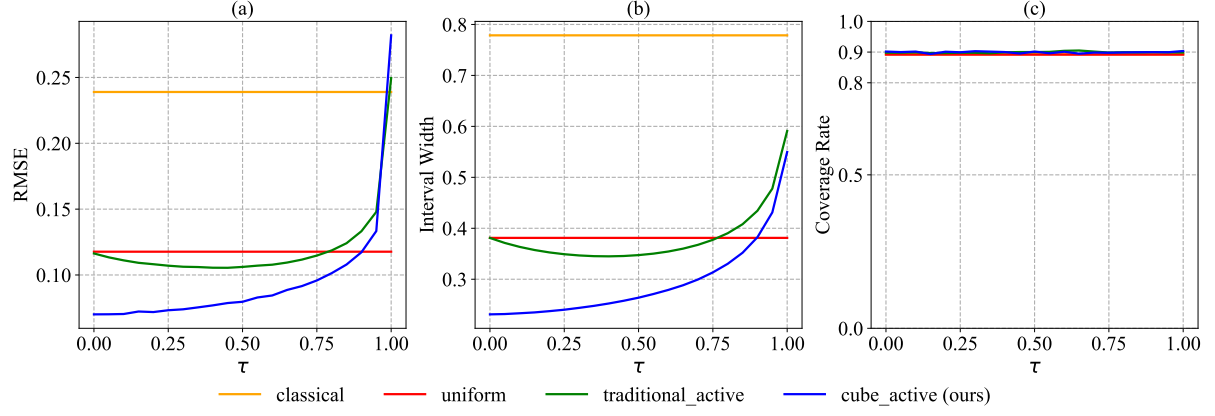


Figure S11: Performance comparison on the synthetic Nonlinear dataset across different τ values while keeping budget=0.1 fixed. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

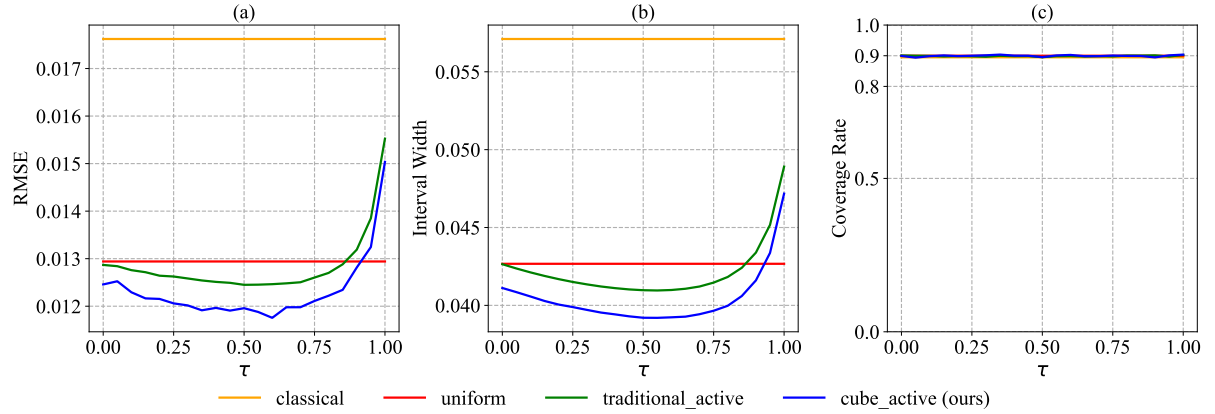


Figure S12: Performance comparison on the Post-election Survey dataset across different τ values while keeping budget=0.1 fixed. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

S6 Discussion of Uncertainty Quantifications

We have conducted additional experiments to evaluate the effectiveness of different uncertainty quantification (UQ) methods, with the figure illustrating the results of using ensemble variance as a measure of uncertainty [14] on the Bike Sharing dataset (Figure S13) and prediction interval (5%–95% quantiles) from quantile regression [15] on the Nonlinear dataset (Figure S14); both methods demonstrate improved performance compared to the baseline, yet they are inferior to absolute error.

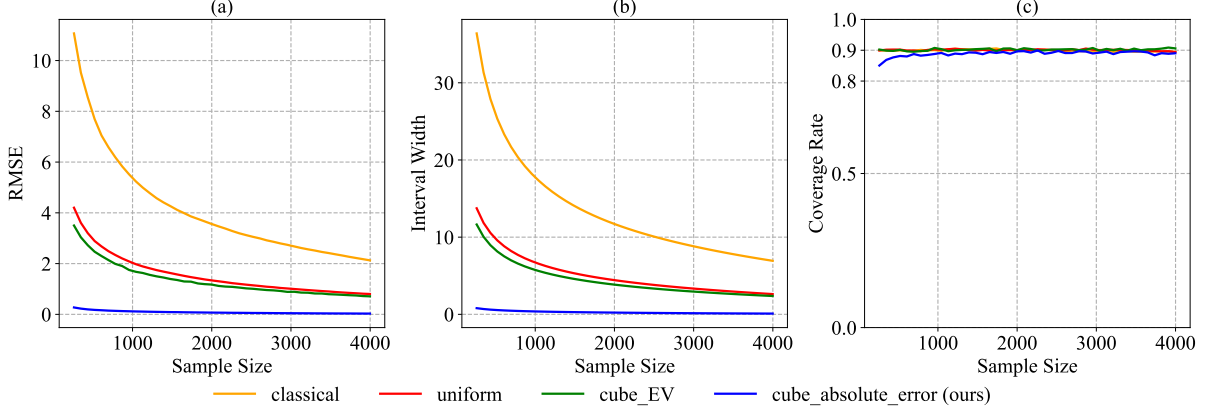


Figure S13: Performance comparison on the Post-election Survey dataset using ensemble variance as a measure of uncertainty. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

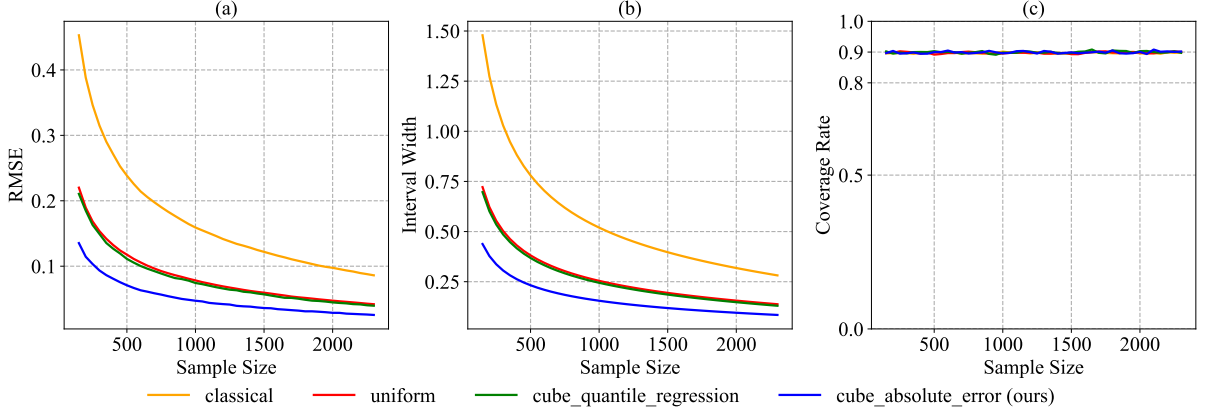


Figure S14: Performance comparison on the Post-election Survey dataset using prediction interval (5%–95% quantiles) from quantile regression. Subfigures depict (a) RMSE, (b) 90% confidence interval width, and (c) empirical coverage rate across different sampling methods.

S7 Coverage Rates for Different Confidence Levels

Across multiple datasets, we have calculated the coverage rate under varying confidence levels, and the results consistently demonstrate that our method remains close to the preset confidence level in all cases. The results presented in this section range from Figure S15 to Figure S24.

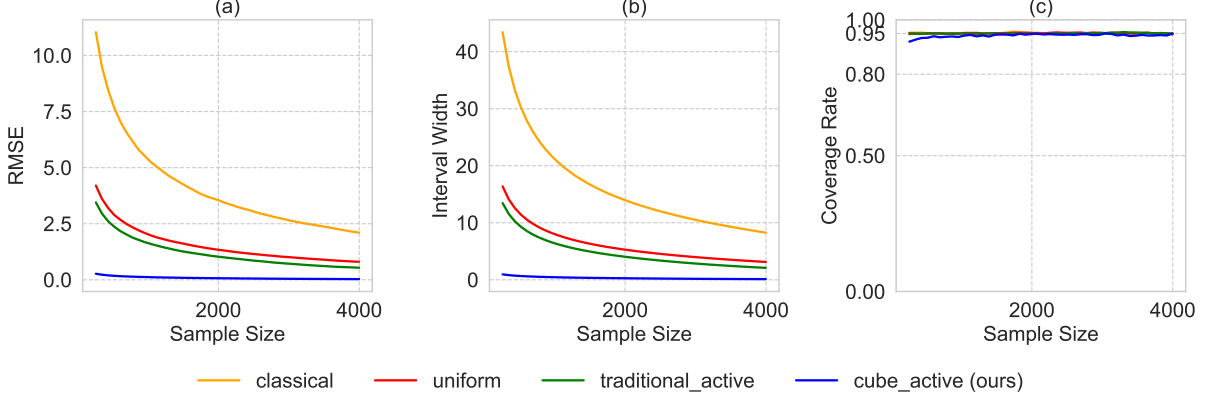


Figure S15: Performance comparison on the Bike Sharing dataset. Subfigures depict (a) RMSE, (b) 95% confidence interval width, and (c) empirical coverage rate across different sampling methods.

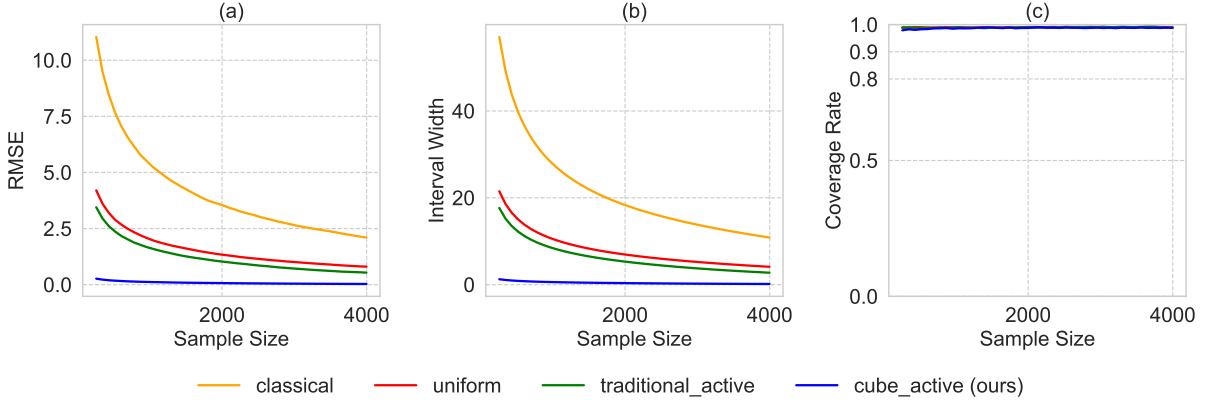


Figure S16: Performance comparison on the Bike Sharing dataset. Subfigures depict (a) RMSE, (b) 99% confidence interval width, and (c) empirical coverage rate across different sampling methods.

S8 Limitations and Future Directions

Our work presents several limitations that warrant discussion. First, when the predictive model achieves exceptionally high accuracy, the differences between our method, prediction-powered inference, and traditional active inference diminish, as all approaches converge to similar efficiency bounds. While this scenario validates the robustness of our framework, it may also increase susceptibility to overfitting, potentially reducing model adaptability under distribution shifts. Second, the computation of confidence intervals in our method requires solving nonlinear equations, which introduces higher time and space complexity compared to closed-form alternatives. However, in practical applications where precision is prioritized over computational speed, such as medical diagnostics or policy planning, this trade-off remains acceptable given sufficient resources.

Future research directions are abundant. The extension of balanced active inference to sequential settings poses intriguing methodological challenges, particularly in maintaining dynamic balancing constraints as new data arrives and model uncertainties evolve. Theoretical guarantees for such sequential adaptations, including martingale-based asymptotic analyses, remain an open problem. Another promising direction

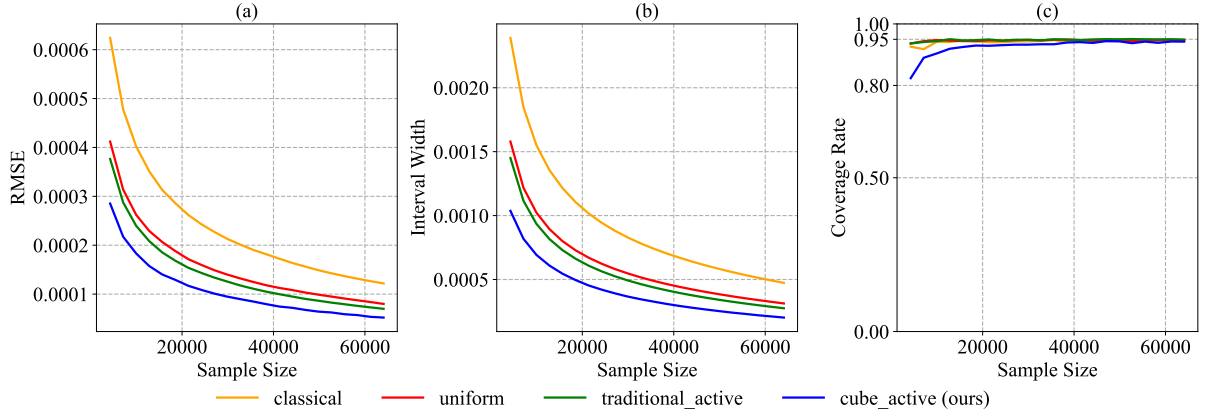


Figure S17: Performance comparison on the Credit Card Fraud Detection dataset. Subfigures depict (a) RMSE, (b) 95% confidence interval width, and (c) empirical coverage rate across different sampling methods.

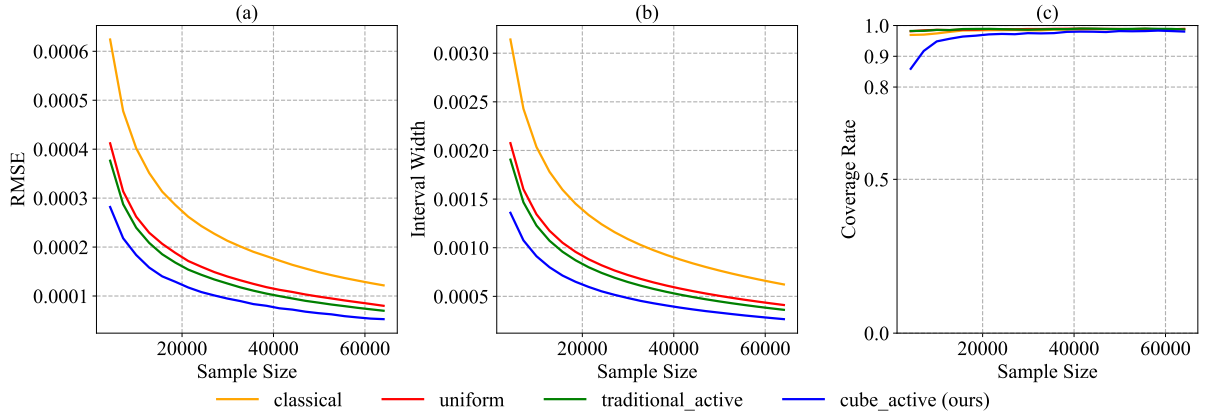


Figure S18: Performance comparison on the Credit Card Fraud Detection dataset. Subfigures depict (a) RMSE, (b) 99% confidence interval width, and (c) empirical coverage rate across different sampling methods.

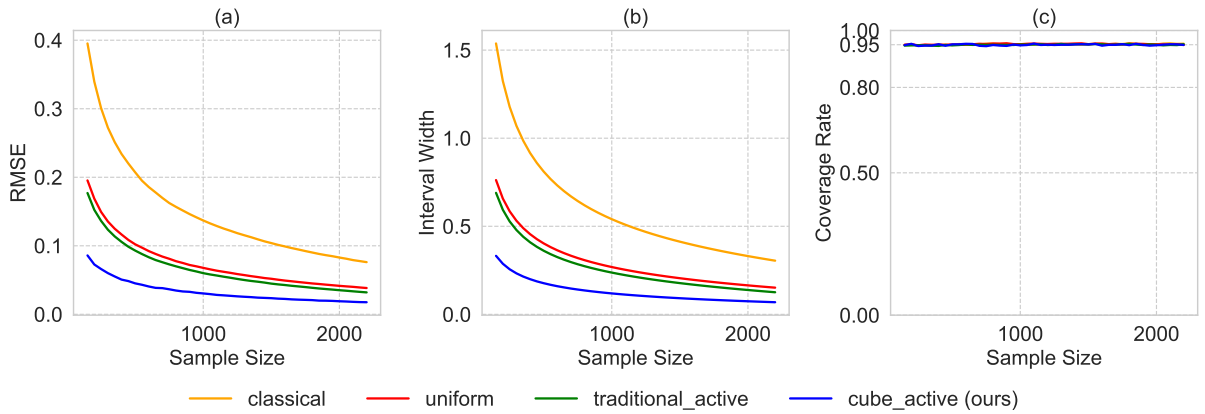


Figure S19: Performance comparison on the Friedman dataset. Subfigures depict (a) RMSE, (b) 95% confidence interval width, and (c) empirical coverage rate across different sampling methods.

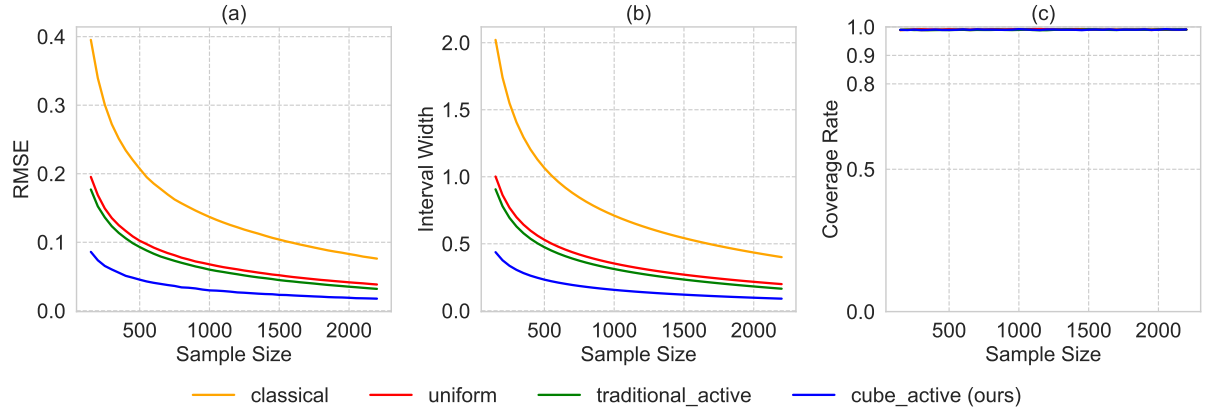


Figure S20: Performance comparison on the Friedman dataset. Subfigures depict (a) RMSE, (b) 99% confidence interval width, and (c) empirical coverage rate across different sampling methods.

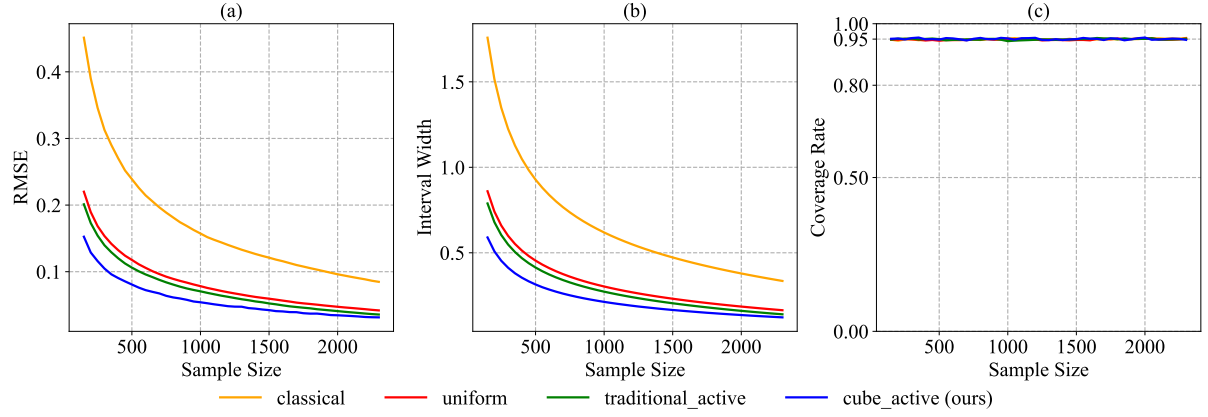


Figure S21: Performance comparison on the Nonlinear dataset. Subfigures depict (a) RMSE, (b) 95% confidence interval width, and (c) empirical coverage rate across different sampling methods.

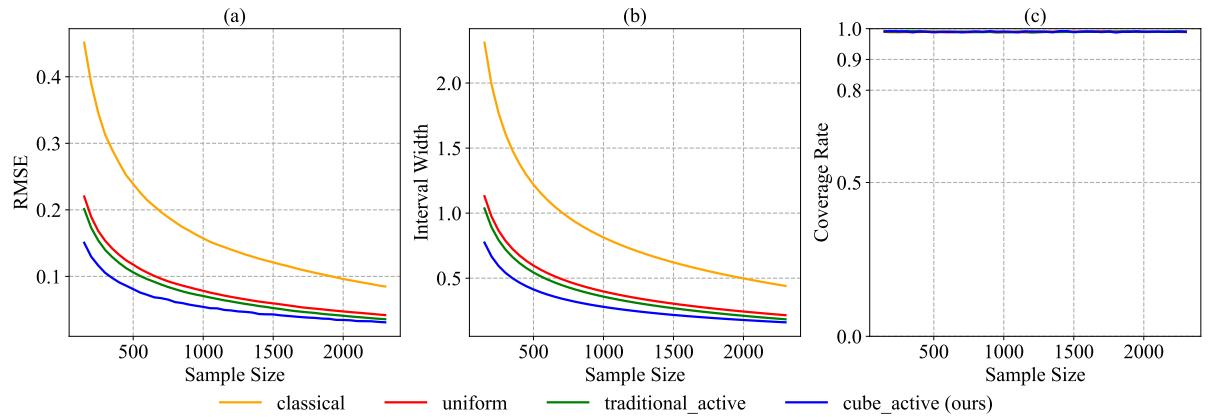


Figure S22: Performance comparison on the Nonlinear dataset. Subfigures depict (a) RMSE, (b) 99% confidence interval width, and (c) empirical coverage rate across different sampling methods.

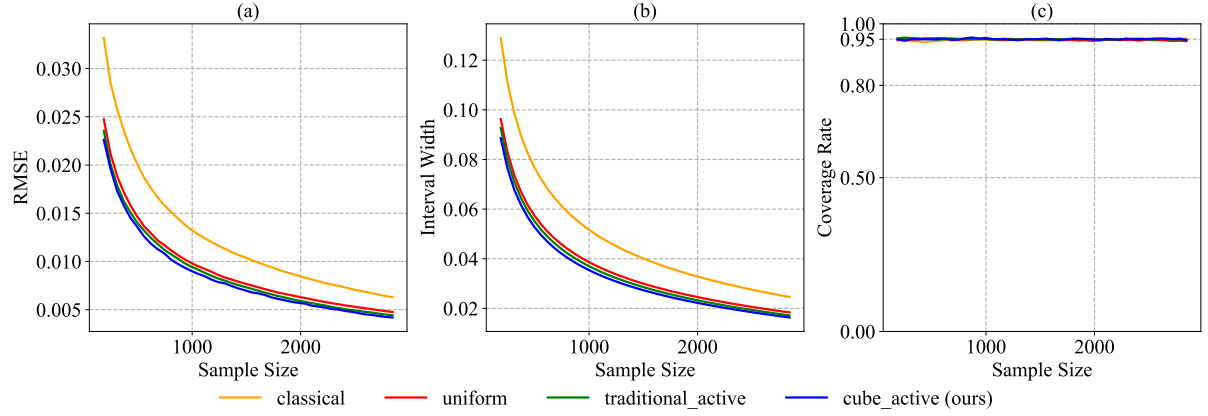


Figure S23: Performance comparison on the Post-election Survey dataset. Subfigures depict (a) RMSE, (b) 95% confidence interval width, and (c) empirical coverage rate across different sampling methods.

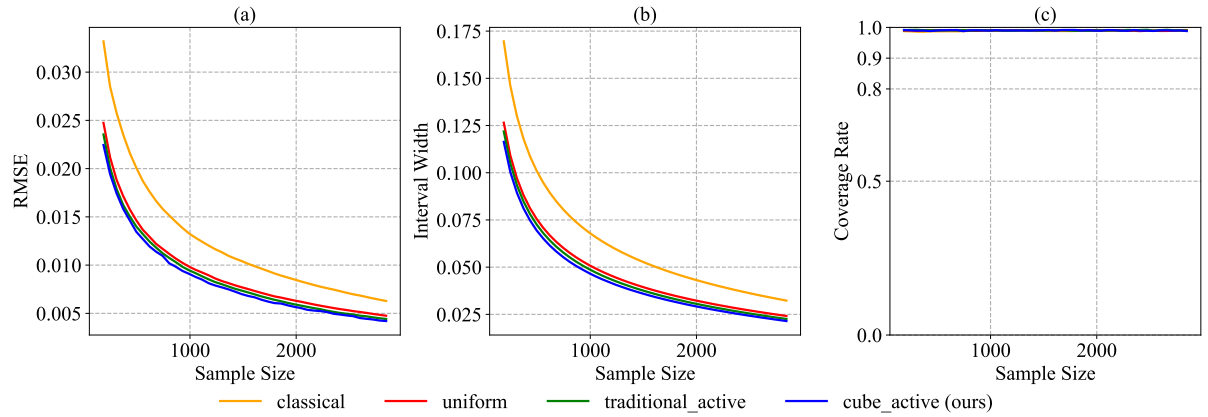


Figure S24: Performance comparison on the Post-election Survey dataset. Subfigures depict (a) RMSE, (b) 99% confidence interval width, and (c) empirical coverage rate across different sampling methods.

involves generalizing our framework to M-estimation problems, where the balancing constraints could be applied to influence functions or other statistical functionals. Establishing theoretical properties for these generalizations, such as consistency and asymptotic normality, would broaden the applicability of our approach.

S9 Societal Impact

Our work introduces a balanced active inference framework that significantly reduces labeling costs for data-driven research in domains such as medical analysis, remote sensing, and population census. By improving statistical efficiency through uncertainty-aware balanced sampling, this method enables resource-constrained organizations to conduct large-scale surveys and studies with minimal annotation budgets, potentially democratizing access to high-quality data analysis. This could accelerate progress in public health initiatives, environmental monitoring, and evidence-based policymaking by making data collection more affordable and scalable.

However, several societal considerations warrant attention. First, while our approach mitigates selection bias through balancing constraints, imperfect uncertainty estimation or systematic biases in training data could still propagate disparities in downstream decisions, particularly if deployed in fairness-critical applications like social benefit allocation or criminal justice risk assessments. Second, the efficiency gains from active sampling might incentivize excessive data collection from vulnerable populations if ethical safeguards are not implemented. Third, although the method itself does not directly handle sensitive information, its application to surveys involving personal data (e.g., census or healthcare records) necessitates stringent privacy protections to prevent re-identification risks when combining actively sampled labels with auxiliary features.

We recommend deploying this methodology with explicit fairness audits, differential privacy mechanisms, and transparency about sampling criteria to prevent misuse. Future extensions could integrate ethical constraints directly into the balancing framework to align statistical efficiency with equity objectives.

References

- [1] Jean-Claude Deville and Yves Tillé. Efficient balanced sampling: The cube method. *Biometrika*, 91(4):893–912, 2004.
- [2] Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 12 1976.
- [3] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.
- [4] Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 62993–63010. PMLR, 21–27 Jul 2024.
- [5] Yves Tillé. Ten years of balanced sampling with the cube method: An appraisal. *Survey methodology*, 37(2):215–226, 2011.
- [6] Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.
- [7] Michael Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: <https://doi.org/10.24432/C53W3X>.
- [8] Kolby Nottingham Markelle Kelly, Rachel Longjohn. The uci machine learning repository, 2024.
- [9] I-C Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28(12):1797–1808, 1998.
- [10] Abhinav Kumar. Life expectancy (WHO). Kaggle, 2020. <https://www.kaggle.com/datasets/kumajarshi/life-expectancy-who>.

- [11] Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.
- [12] Kam Hamidieh. Superconductivity Data. UCI Machine Learning Repository, 2018. DOI: <https://doi.org/10.24432/C53P47>.
- [13] Pew. American trends panel (atp) wave 79, 2020. <https://www.pewresearch.org/science/dataset/american-trends-panel-wave-79/>.
- [14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- [15] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32, 2019.